Analysis of bulk RNA-seq data - Part II: From counts to DGE

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2CUdS9z[1]

February 26, 2019

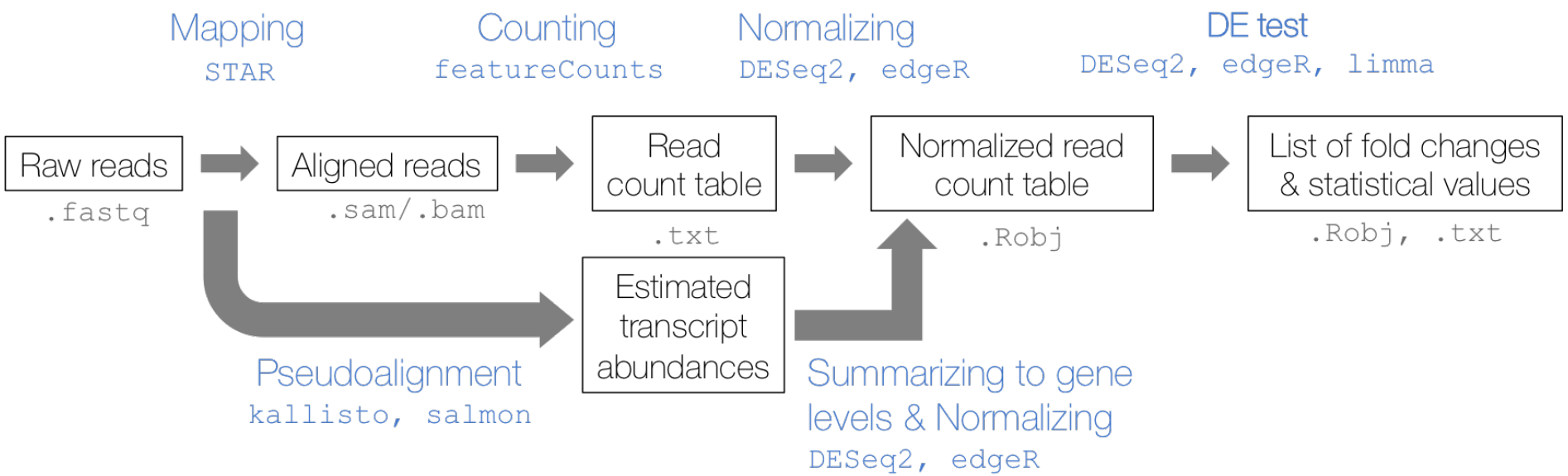
[1] http://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2018/

1. Normalization of read counts
2. Exploratory analyses
3. Differential gene expression
4. Downstream analyses
5. References

Many slides today were influenced or taken from the excellent book **Data Analysis for the Life Sciences** by Rafael Irizarry and Michael Love, and training material developed by the **Harvard Chan Bioinformatics Core**.

Go and check them out for even more details! The Harvard Chan Bioinformatics Core's material can be found at their github page: https://github.com/hbctraining/DGE_workshop

General bioinformatics workflow for RNA-seq data



Normalization of read counts

Read counts are influenced by numerous factors, not just expression strength

Raw counts[2]: number of reads (or fragments) overlapping with the union of exons of a gene.

> Raw count numbers are not just a reflection of the actual number of captured transcripts!

They are strongly influenced by:

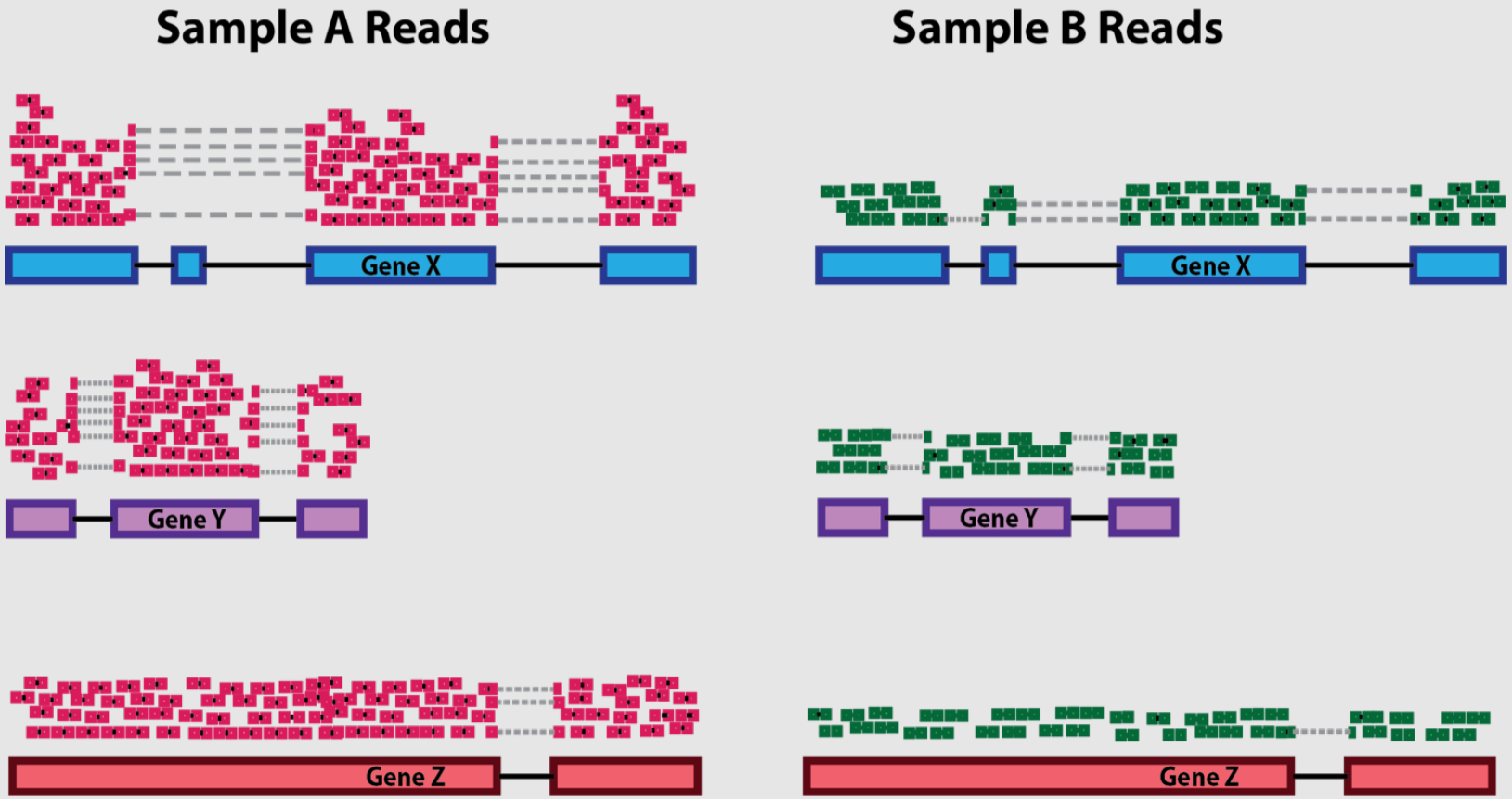
- sequencing depth
- gene length
- DNA sequence content (% GC)
- expression of all other genes in the same sample

[2] also true for "estimated" gene counts from pseudoaligners

Influences on read count numbers

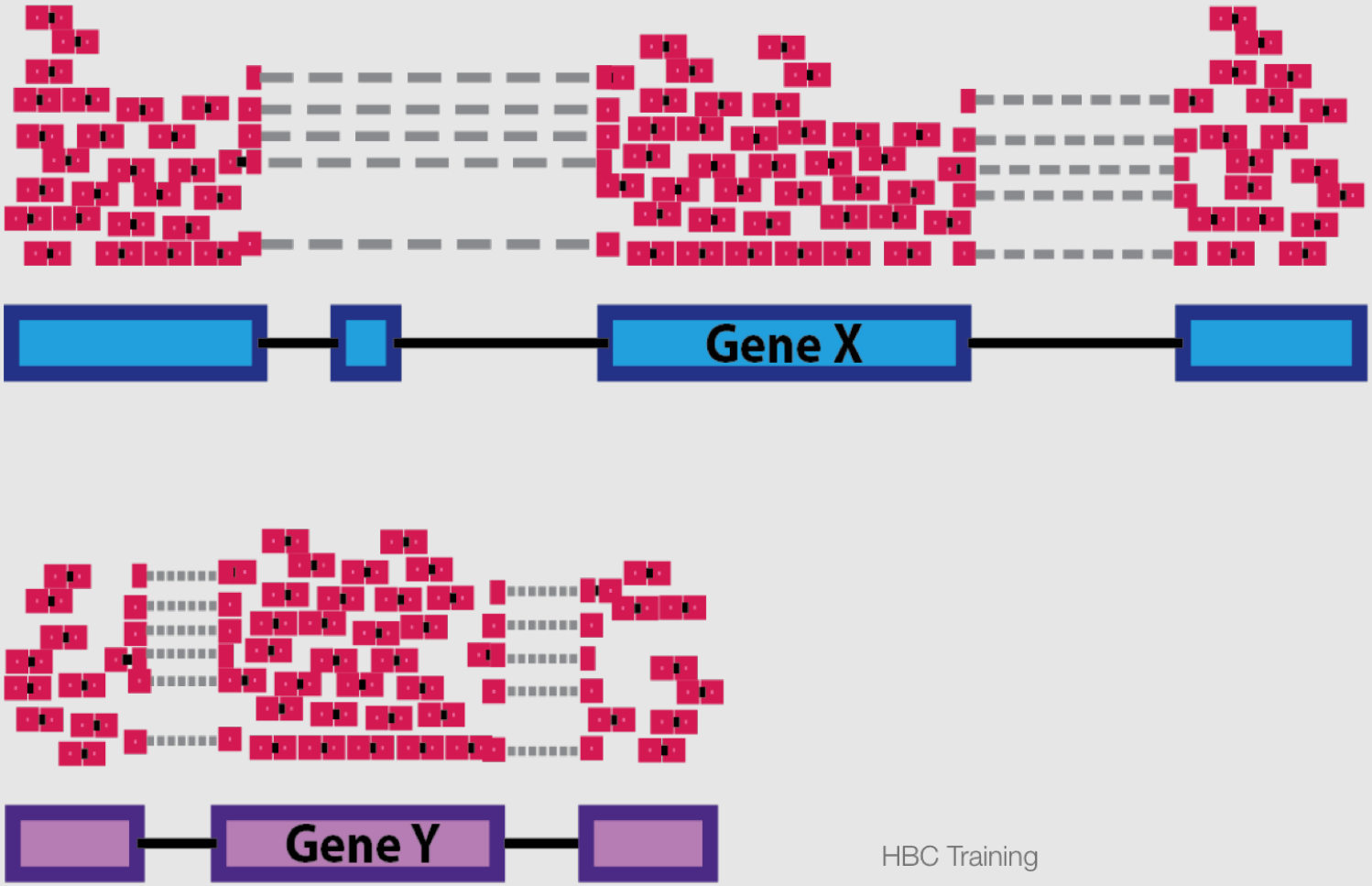
1. Sequencing depth

sequencing depth of Sample A $\gg$ Sample B



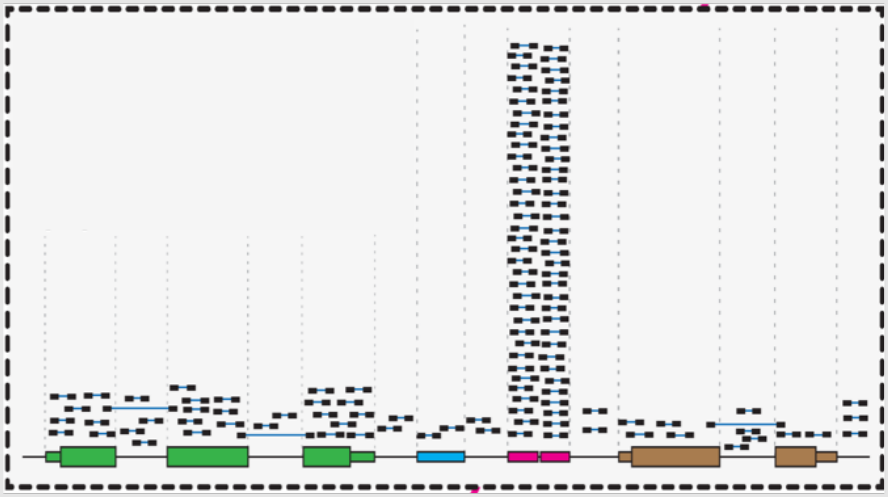
Influences on read count numbers

2. Gene length (and GC bias)

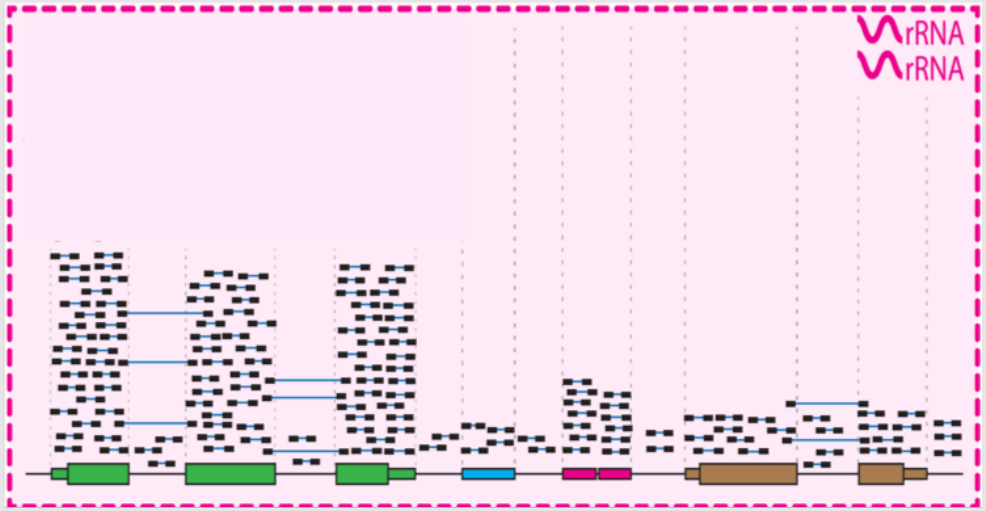


Influences on read count numbers

3. RNA composition - individual gene abundances



very highly expressed transcript soaks up significant portion of the reads reducing the range of read counts available for other transcripts



in the absence of that highly expressed transcript, the remaining transcripts' expression differences become more clear

Influences on read count numbers - summary

- gene length
- transcript sequence (% GC)

need to be corrected when comparing different **genes**

- sequencing depth
- expression of all other genes within the same sample

need to corrected when comparing the same gene between different **samples**

Which biases are relevant for comparing different samples?

Different units for expression values

- **Raw counts**: number of reads/ fragments overlapping with the union of exons of a gene

$$X_i$$

- **[RF]PKM**: Reads/Fragments per Kilobase of gene per Million reads mapped – AVOID!

$$RPKM_i = \frac{X_i}{(\frac{l_i}{10^3})(\frac{N}{10^6})}$$

gene length seq. depth

- **TPM**: Transcripts Per Million

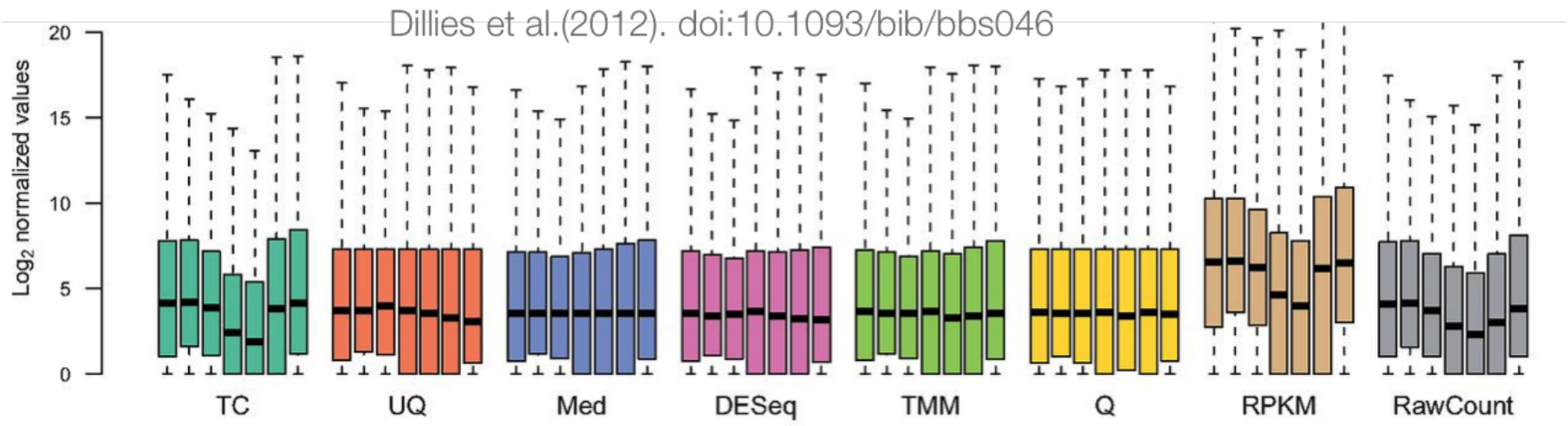
$$TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}} * 10^6$$

gene read counts per bp

all gene counts over *all* gene bp

- **rlog**: log2-transformed count data normalized for small counts and library size (DESeq2)

Why not RPKMs?



- [RF]PKM values are not comparable between samples – Do NOT use them!
- if you need normalized expression values for exploratory plots, use TPM or DESeq2's rlog values

Working with read counts

- Download the featureCounts results to your laptop.
- Read the featureCounts results into R.
- Let's normalize!

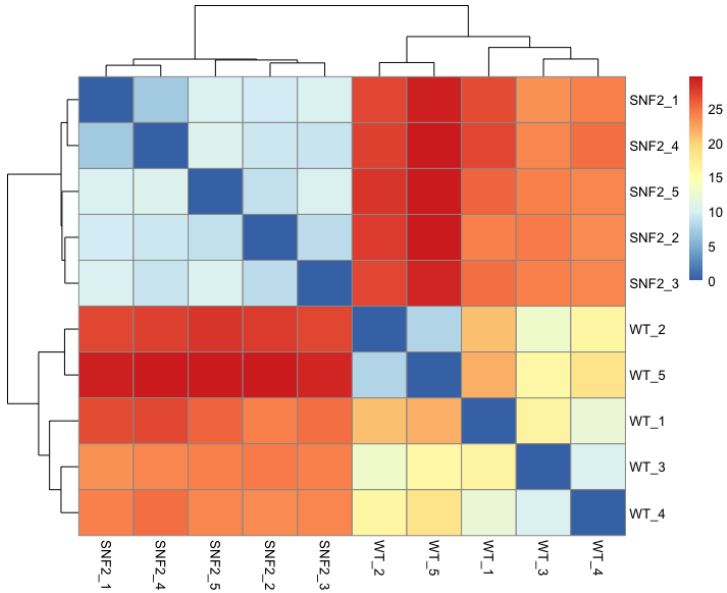
Exploratory analyses

Exploratory analyses

Exploratory analyses **do not test a null hypothesis**! They are meant to familiarize yourself with the data to discover biases and unexpected variability!

Typical exploratory analyses:

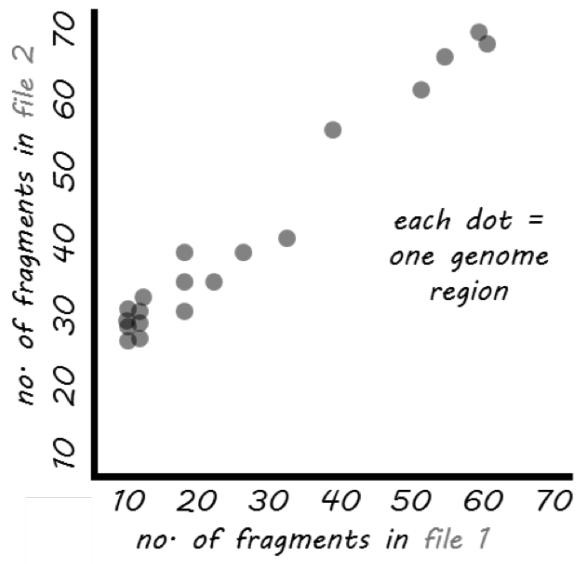
- **correlation** of gene expression between different samples
- (hierarchical) **clustering**
- **dimensionality reduction** methods, e.g. PCA
- dot plots/**box plots**/violin plots of individual genes

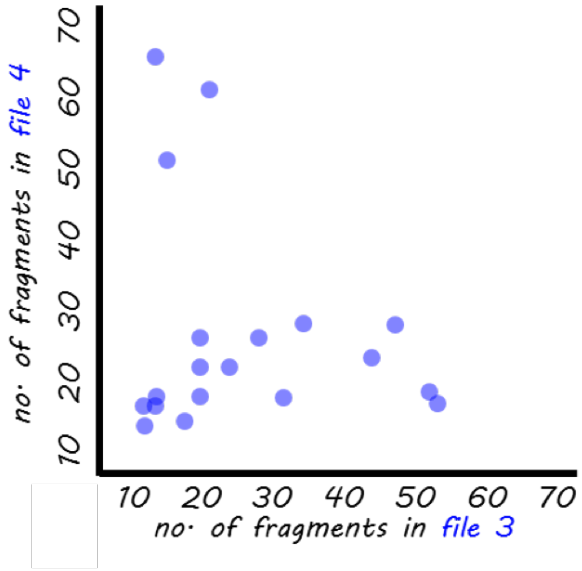


Use **normalized and transformed** read counts for data exploration!

Pairwise correlation of gene expression values

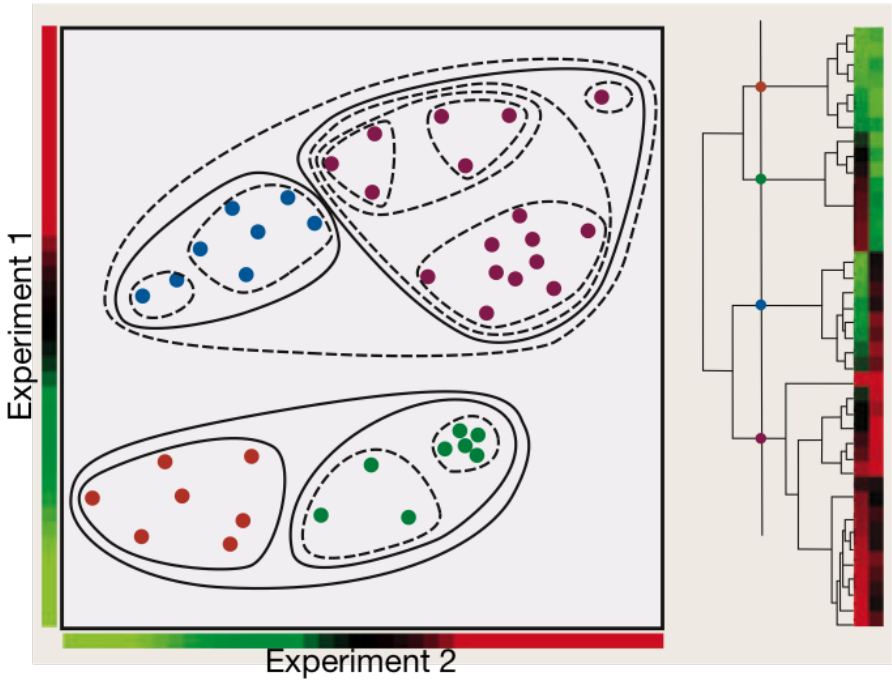
- replicates of the same condition should show high correlations (>0.9)
- **Pearson** method: *metric* differences between samples
  - influenced by outliers
  - covariance of two variables divided by the product of their standard deviation
  - suitable for normally distributed values
- **Spearman** method: based on *rankings*
  - less sensitive
  - less driven by outliers
- R function: `cor()`





Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

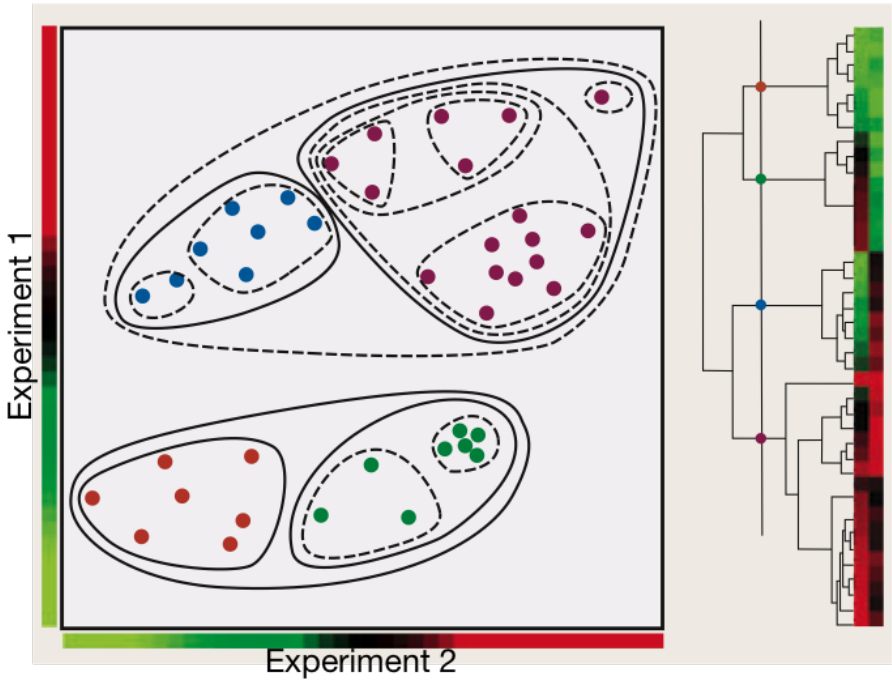


single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

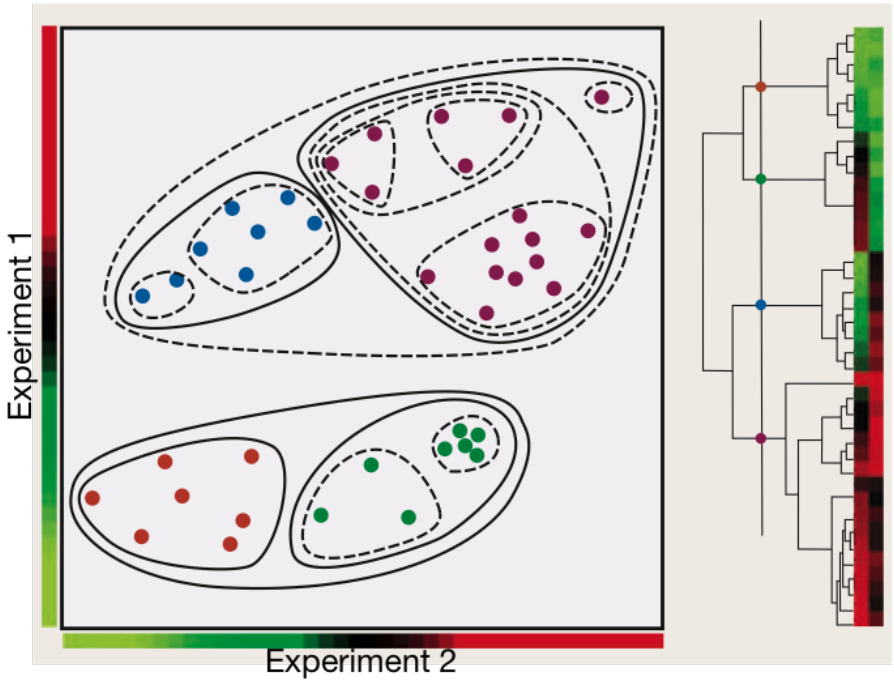


single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

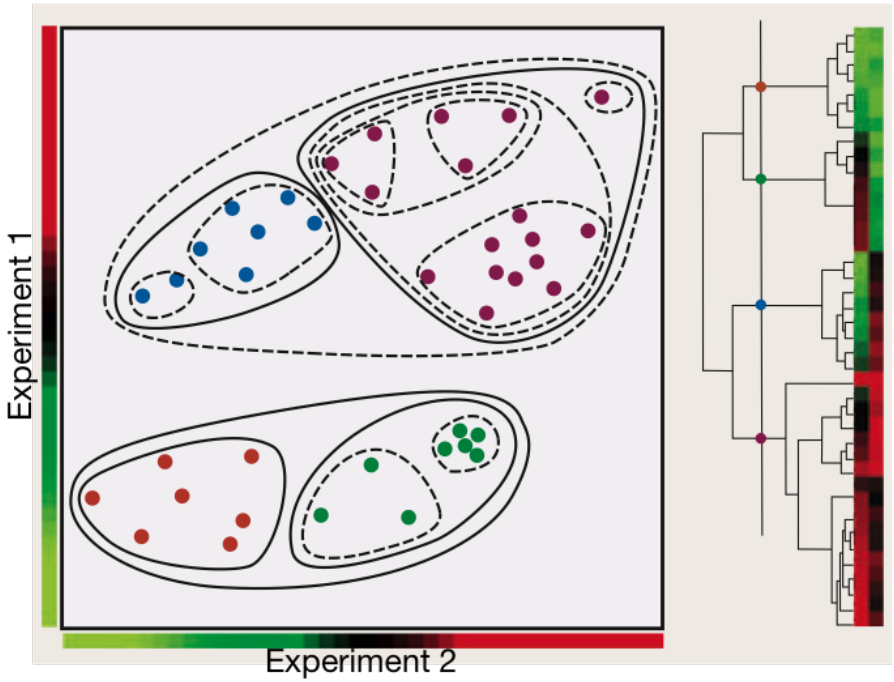


single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

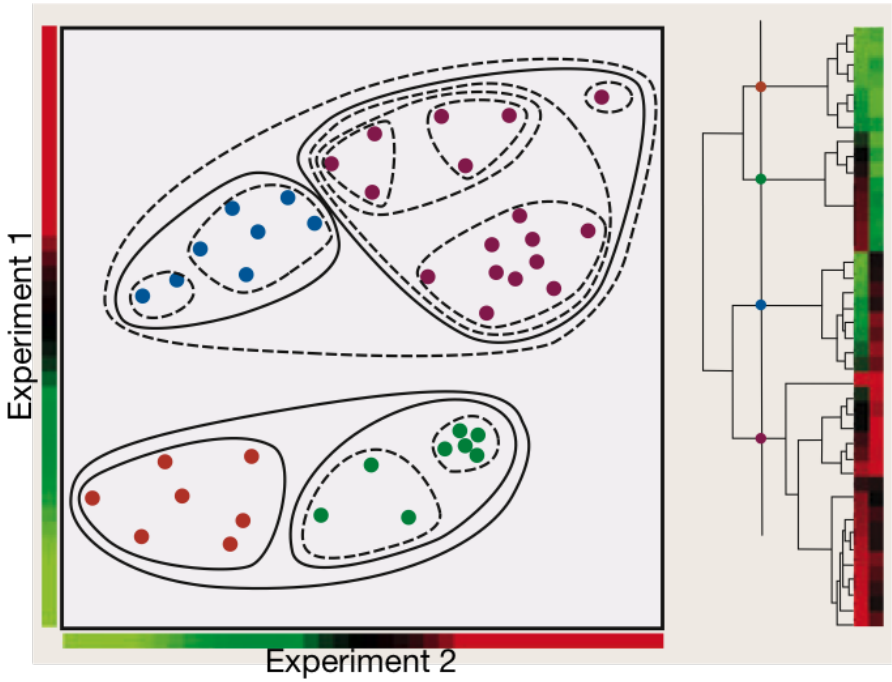


single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

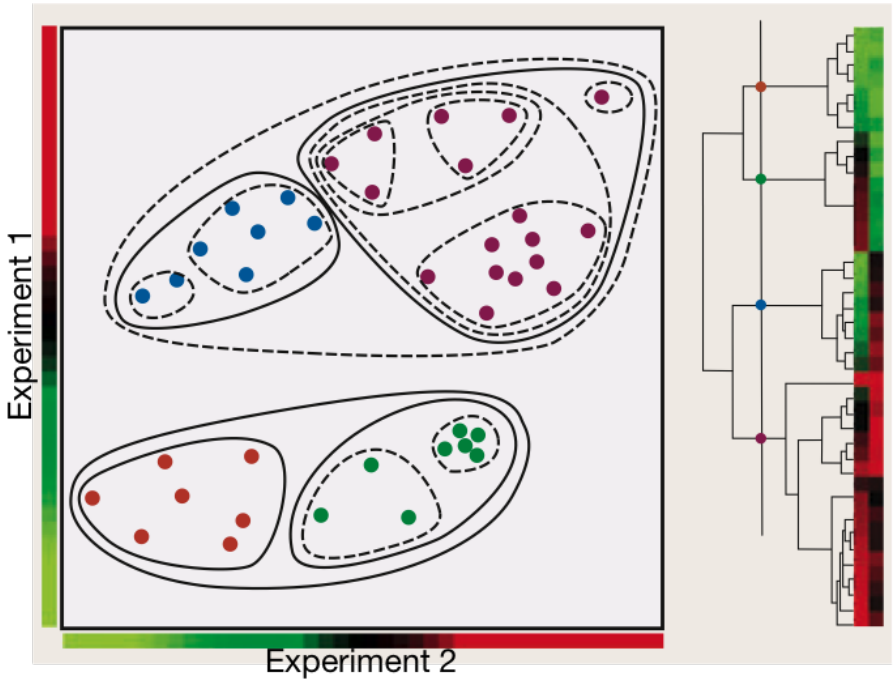


single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

Hierarchical clustering – grouping similar samples

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.



single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

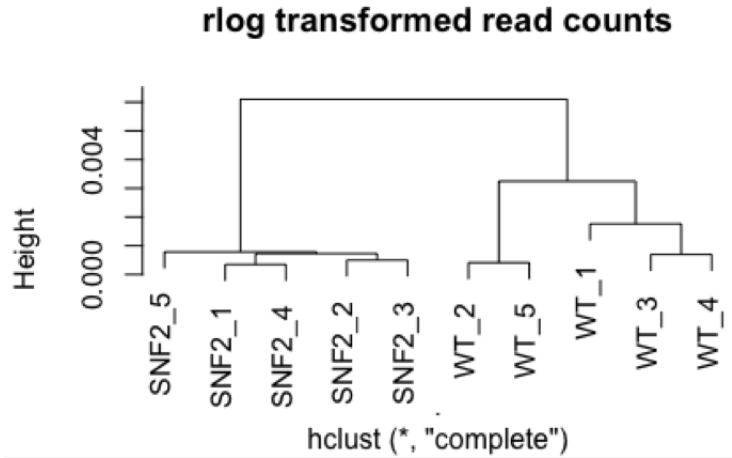
- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

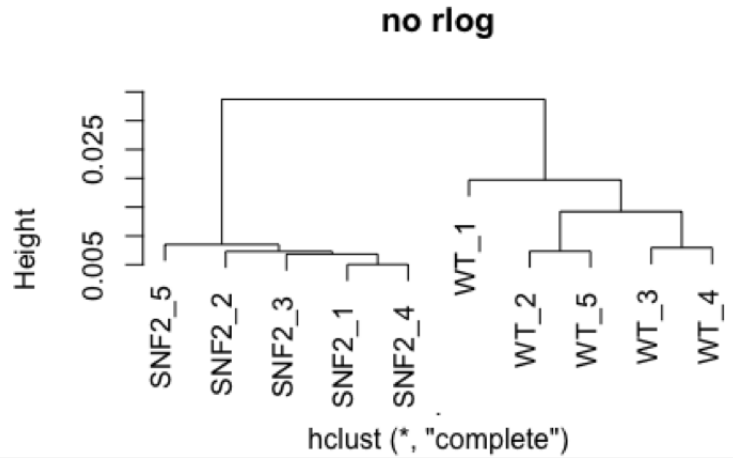
Hiearchical clustering - R code

```
## calculate the correlation between columns of a matrix
pw_cor <- cor(rlog.norm.counts, method = "pearson" )

## use the correlation as a distance measure
distance.m_rlog <- as.dist(1 - pw_cor)

##  plot() can directly interpret the output of hclust() to generate
## a dendrogram
plot( hclust(distance.m_rlog),
      labels = colnames(rlog.norm.counts),
      main = "rlog transformed read counts")
```





Principal component analysis – capturing variability

Goal: reduce the dataset to have fewer dimensions, yet approx. preserve the distance between samples

starting point: matrix with expression values per gene and sample, e.g. 6,600 genes x 10 samples

	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5	WT_1	WT_2	WT_3	WT_4	WT_5
YDL248W	109	84	100	112	62	47	65	60	95	43
YDL247W.A	0	1	1	0	3	0	0	1	0	0
YDL247W	6	6	1	3	4	2	3	4	7	9
YDL246C	6	6	1	4	4	1	3	2	4	0
YDL245C	1	6	9	5	3	6	2	5	5	6
YDL244W	79	59	49	60	37	9	8	12	30	14

```
assay(DESeq.rlog)[topVarGenes,])
%>% t %>% prcomp
```

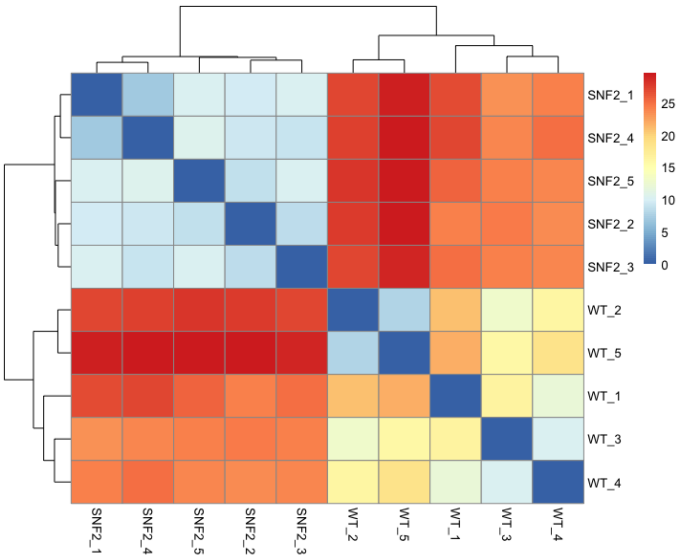
transformed into 6,600 **principal components** x 10 samples

	PC1	PC2
SNF2_1	-9.322866	0.8929154
SNF2_2	-9.390920	-0.6478100
SNF2_3	-9.176814	0.3460428
SNF2_4	-9.693035	1.2174519
SNF2_5	-9.450847	-0.3668670
WT_1	8.378671	-6.3321623
WT_2	10.421518	4.6749399
WT_3	8.486379	-1.1793146
WT_4	8.517490	-4.5814481

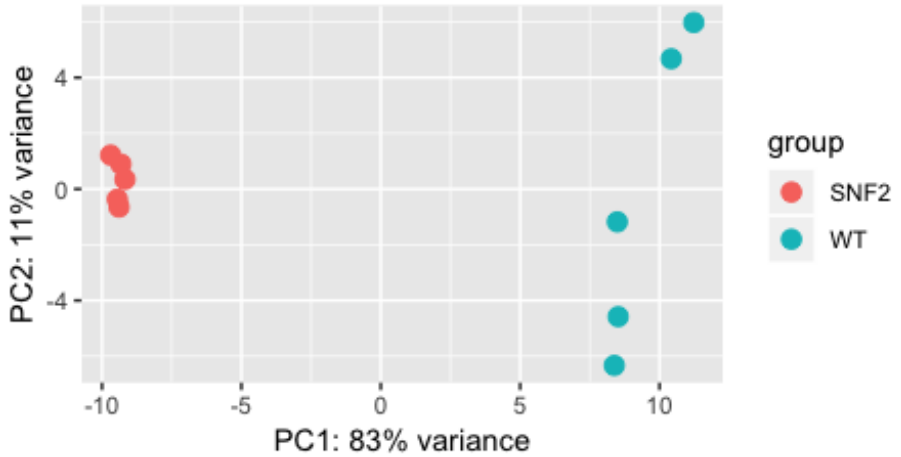
- linear combi of optimally weighted observed variables
- the vectors along which the variation between samples is maximal
- PC1-3 are usually sufficient to capture the major trends!

PCA vs. hierarchical clustering

- often similar results because both techniques should capture the most dominant patterns
- PCA will always be run on just a subset of the data!
- clustering will ALWAYS return clusters, PCA may not if the patterns of variation are too random



See `practical_exploratory.Rmd` R code to generate exploratory plots.
Use the `pcaExplorer` package!



See the chapter "Distance and Dimension Reduction" in Irizarry and Love [2015] for more details and the StatQuest video(s) on youtube.

Differential gene expression

Understand your null hypothesis!

- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) **counts per gene** used by `DESeq2/edgeR/limma`
  - see Soneson et al. [2015] and bioconductor's `tximport` package vignette for details

- **DTU**: Differential **Transcript Usage**
  - Has the **isoform composition** for a given gene changed? I.e. are there different *dominant* isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)
  - see Love et al. [2018] for details

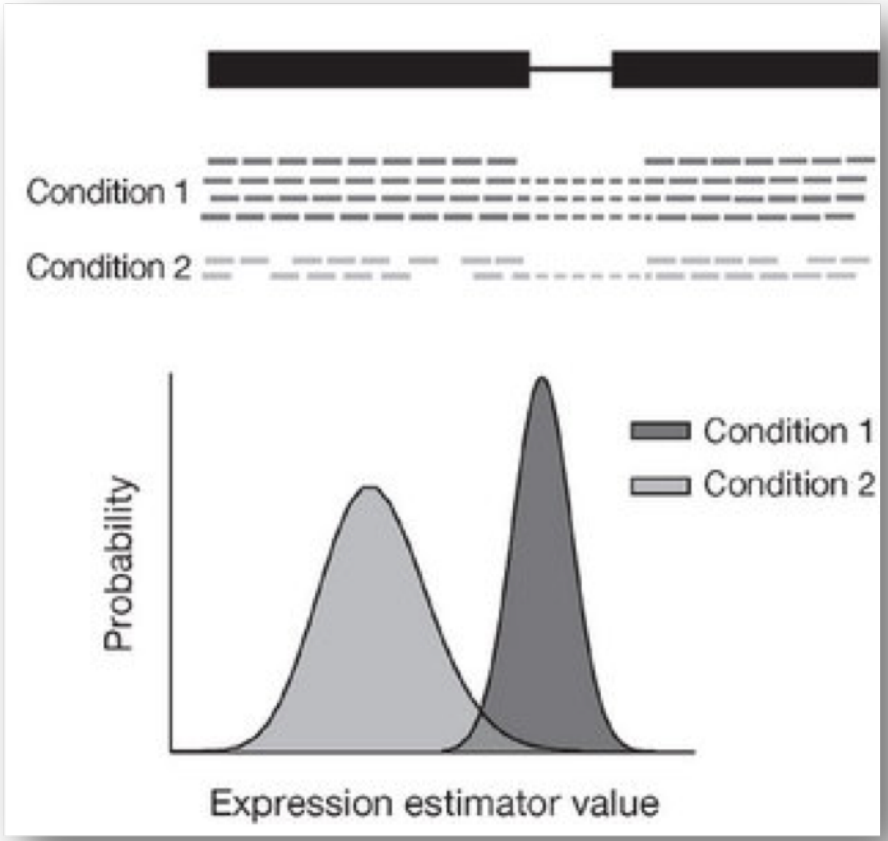
Understand your null hypothesis!

- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) **counts per gene** used by `DESeq2/edgeR/limma`
  - see Soneson et al. [2015] and bioconductor's `tximport` package vignette for details

- **DTU**: Differential **Transcript Usage**
  - Has the **isoform composition** for a given gene changed? I.e. are there different *dominant* isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)
  - see Love et al. [2018] for details

DGE basics

H_0: There is no difference in the read distributions of the 2 conditions.



1 test per gene!

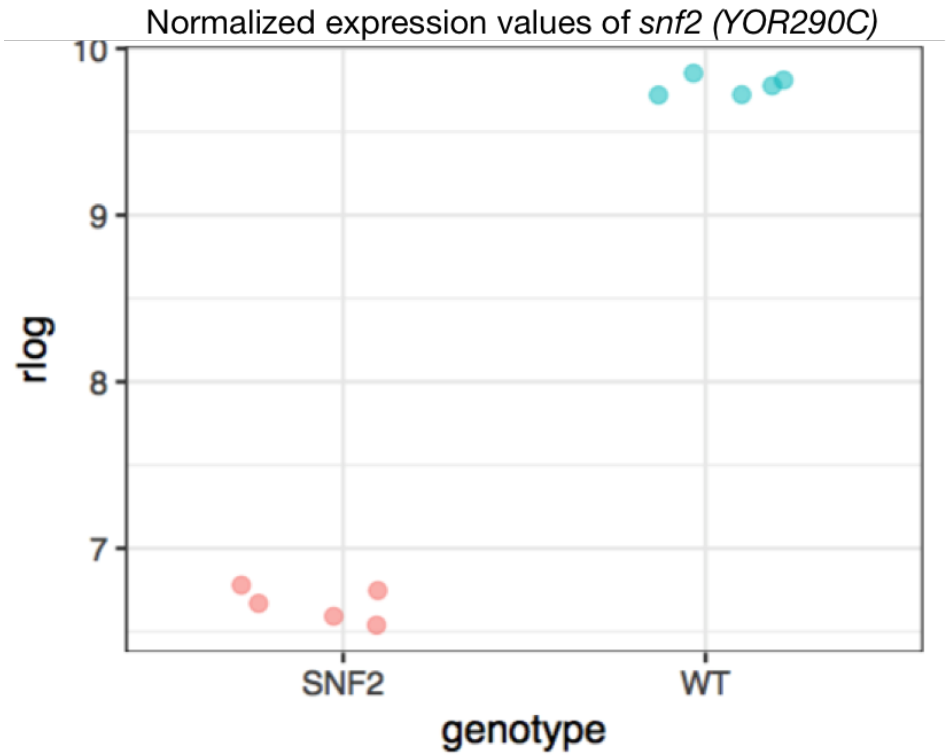
1. Estimate **magnitude** of DE taking into account differences in sequencing depth, technical, and biological read count variability.

logFC

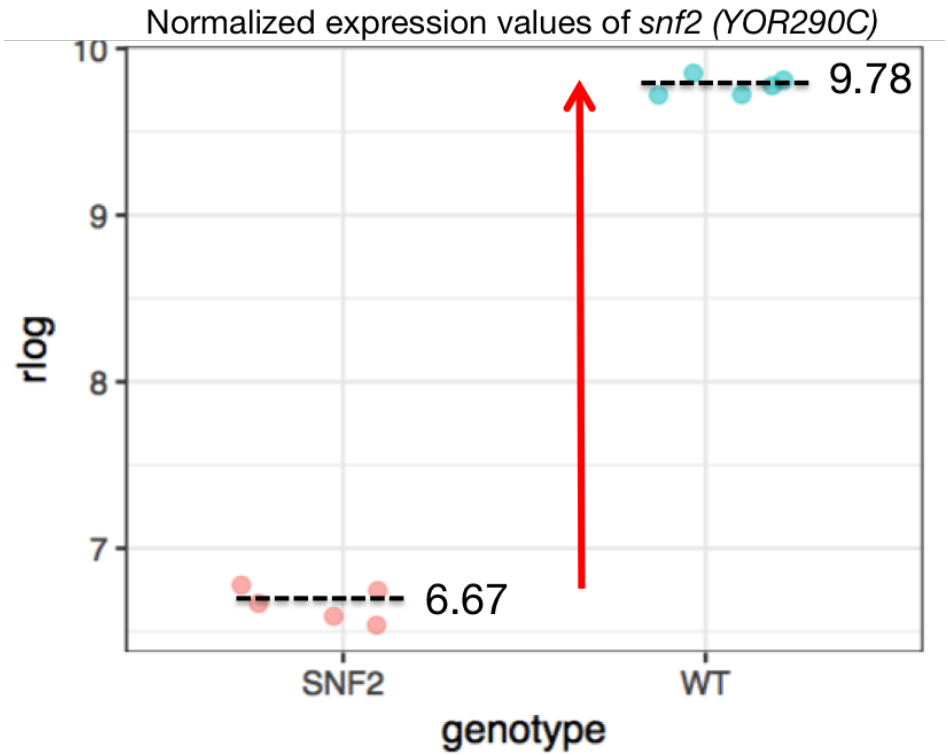
2. Estimate the **significance** of the difference accounting for performing thousands of tests.

(adjusted) p-value

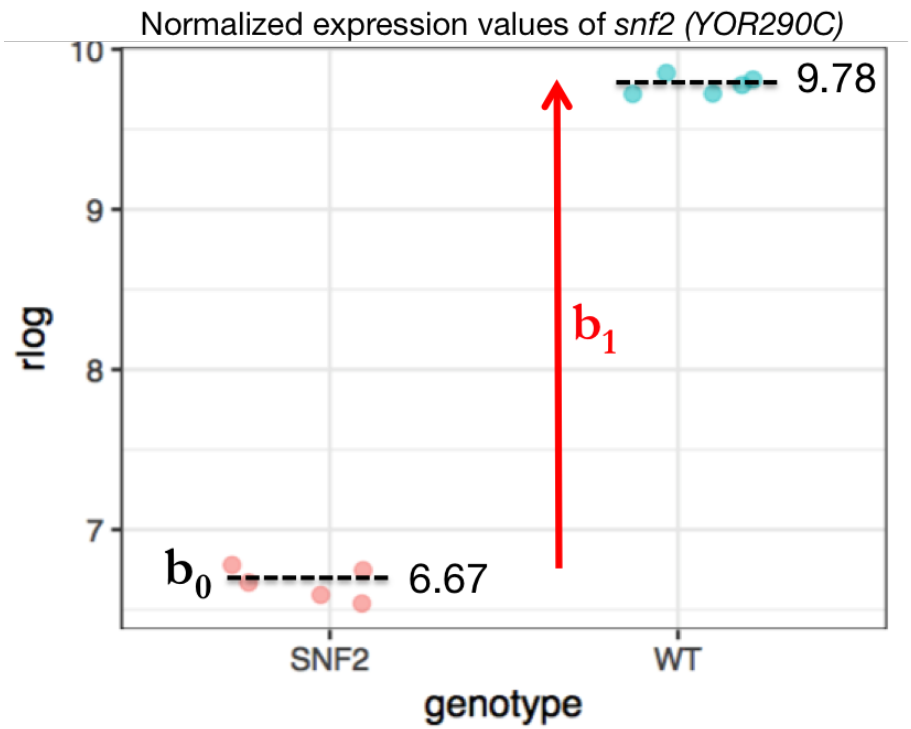
Applying linear models for read count modeling



Applying linear models for read count modeling



Applying linear models for read count modeling

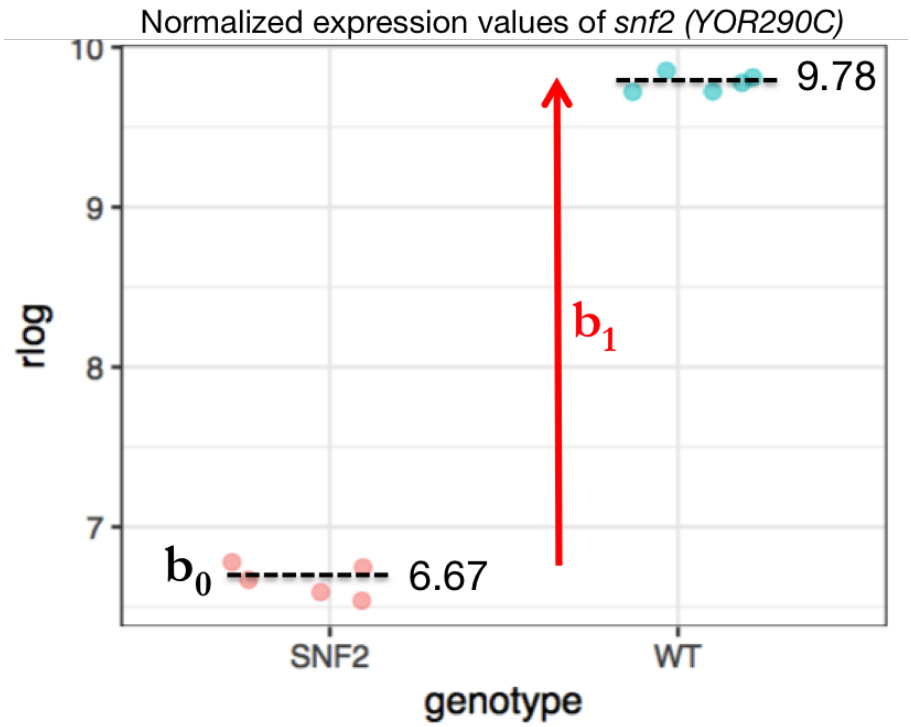


To describe all expression values of one (!) example gene (*snf2*), we can use a linear model like this:

$$\underset{\text{expression values}}{Y} = \underset{\text{intercept}}{b_0} + \underset{\substack{\text{genotype (discrete factor here!)}}}{b_1 * x} + e$$

Linear models model a response variable as a linear combination of predictors (betas), plus randomly distributed noise (e).

Applying linear models for read count modeling



To describe all expression values of one (!) example gene (*snf2*), we can use a linear model like this:

$$\underset{\text{expression values}}{Y} = \underset{\text{intercept}}{b_0} + b_1 * \underset{\substack{\text{genotype (discrete factor here!)}}}{x} + e$$

Linear models model a response variable as a linear combination of predictors (betas), plus randomly distributed noise (e).

- b_0: **intercept**, i.e. average value of the baseline group
- b_1: **difference** between baseline and non-reference group
- x: 0 if genotype == "SNF2", 1 if genotype == "WT"

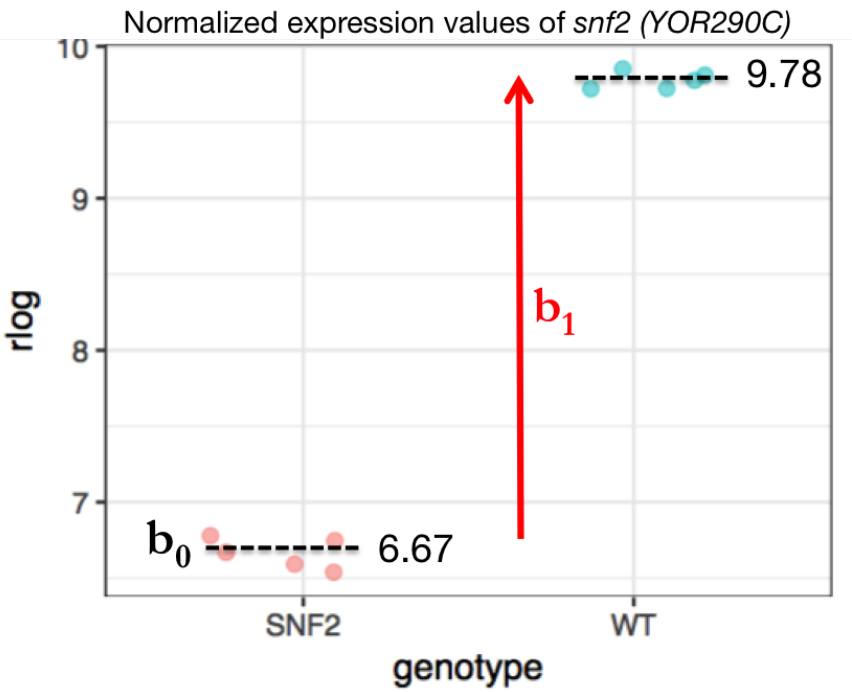
Model formulae syntax in R

- regression functions in R (e.g., `lm()`, `glm()` use a "model formula" interface
- the basic format is: `response variable ~ explanatory variables` [3], e.g. `lm( y ~ x )`

If you find yourself using linear models and somewhat complicated experimental designs more often than not, we strongly recommend to work through **chapters 4 and 5** of the PH525x series **Biomedical Data Science** [Irizarry and Love, 2016]

[3] Tilde means "is modeled by" or "is modeled as a function of". See King [2016] for more details on the specialy meaning of mathematical operators within formula contexts.

Applying linear models for read count modeling



- b_0: **intercept**, i.e. average value of the baseline group
- b_1: **difference** between baseline and non-reference group
- x: 0 if genotype == "SNF2", 1 if genotype == "WT"

Describe expression values *snf2* using a linear model:

$$\underset{\text{expression values}}{Y} = \underset{\text{intercept}}{b_0} + b_1 * \underset{\substack{\text{genotype (discrete factor here!)}}}{x} + e$$

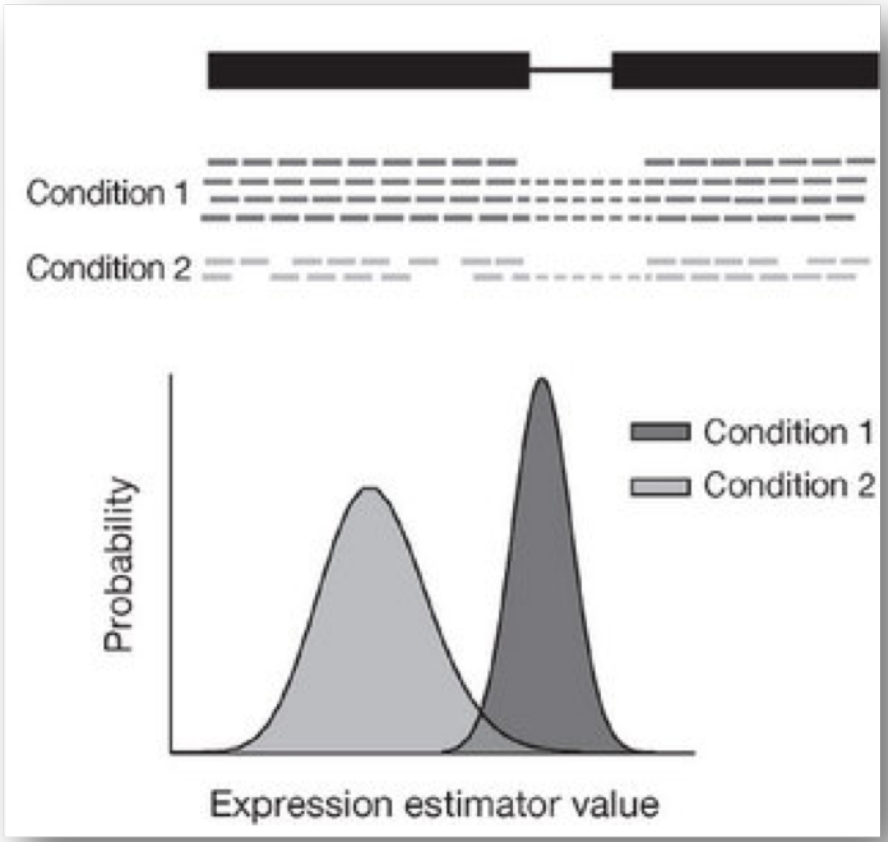
Factor of interest (b_1) can be estimated as follows:

```
# 1. FIT the model
> lmfit <- lm(rlog.norm ~ genotype)
# 2. ESTIMATE the coefficients
> coef(lmfit)
(Intercept)    genotypeWT
      6.666         3.111
```

Both values (b0, b1) are **estimates**! (They're spot-on because the values are so clear and the model is so simple!)

DGE basics

H_0: There is no difference in the read distributions of the 2 conditions.



1 test per gene!

1. Estimate **magnitude** of DE taking into account differences in sequencing depth, technical, and biological read count variability.

logFC

2. Estimate the **significance** of the difference accounting for performing thousands of tests.

(adjusted) p-value

DGE steps (formulae from DESeq2)

1. **Fitting** a sophisticated regression model to the read counts (per gene!)
   - library size factor
   - dispersion estimate using information across multiple genes
   - negative binomial distribution of read counts is assumed

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

2. Estimating coefficients to obtain the difference (**log2FC**)
3. **Test** whether the log2FC is "far away" from zero (remember H0!)

DGE steps (formulae from DESeq2)

1. **Fitting** a more sophisticated model to the read counts (per gene!)
   - library size factor
   - dispersion estimate using information across multiple genes
   - negative binomial distribution of read counts is assumed

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

K_{ij}: read counts for gene i and sample j

α_i: gene-specific dispersion parameter (fitted towards the average dispersion)

2. Estimating coefficients to obtain the difference (**log2FC**)
3. **Test** whether the log2FC is "far away" from zero (remember H0!)

DGE steps (formulae from DESeq2)

1. **Fitting** a more sophisticated model to the read counts (per gene!)
   - library size factor
   - dispersion estimate using information across multiple genes
   - negative binomial distribution of read counts is assumed

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

K_{ij}: read counts for gene i and sample j

μ_{ij}: mean expr.

α_i: gene-specific dispersion parameter (fitted towards the average dispersion)

$$\mu_{ij} = s_j q_{ij}$$

s_j: library size factor

2. Estimating coefficients to obtain the difference (**log2FC**)
3. **Test** whether the log2FC is "far away" from zero (remember H0!)

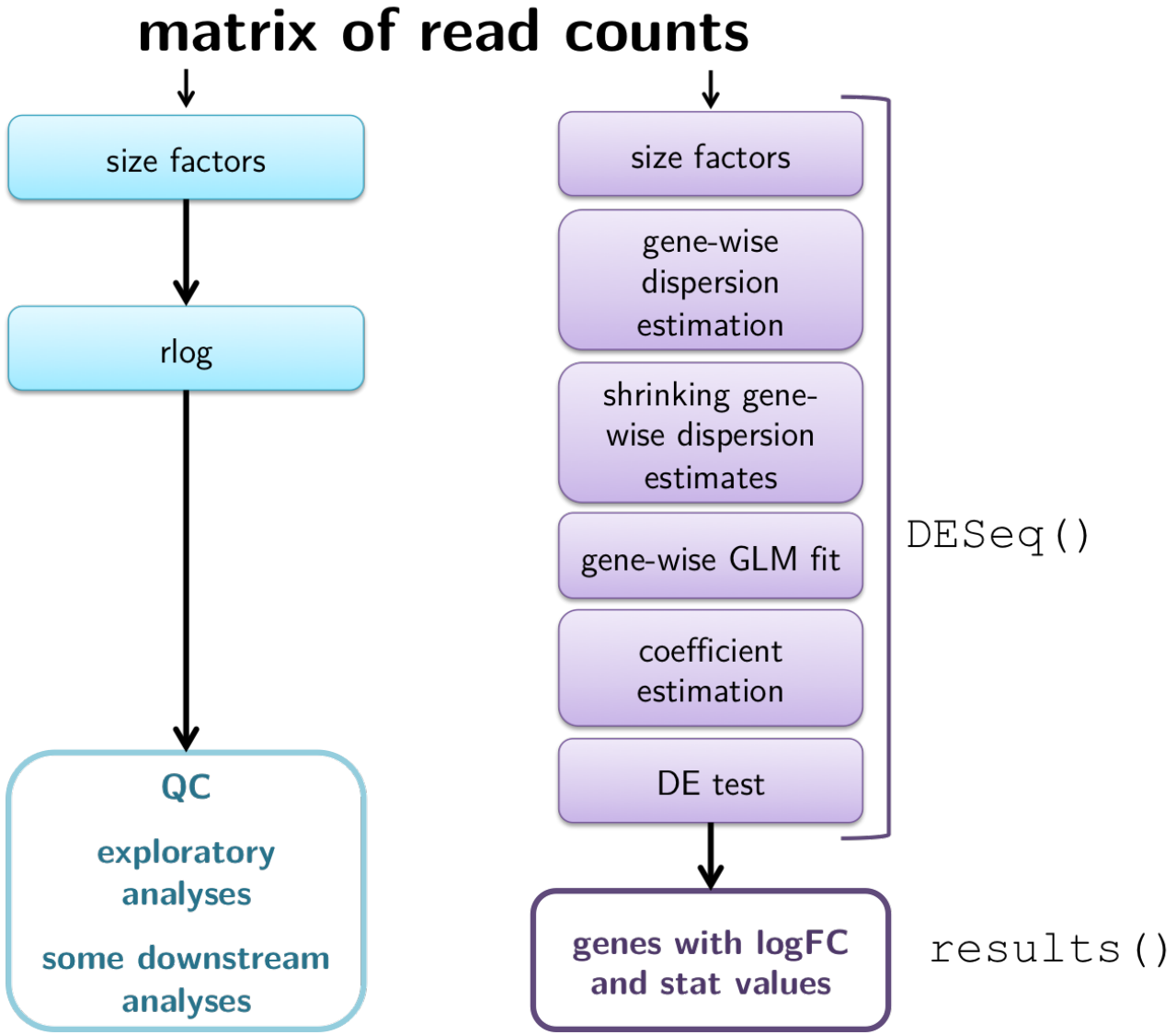
DGE steps (formulae from DESeq2)

1. **Fitting** a more sophisticated model to the read counts (per gene!)
   - library size factor
   - dispersion estimate using information across multiple genes
   - negative binomial distribution of read counts is assumed

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

2. Estimating coefficients to obtain the difference (**log2FC**)
   - define the **contrast of interest**, e.g. `~ condition` or `~ batchEffect + condition`
   - always put the factor of interest last
   - order of the factor levels determines the direction of fold change that is reported
3. **Test** whether the log2FC is "far away" from zero (remember H0!)

Summary: read counts to DGE and other analyses



Comparison of additional tools for DGE analysis

Table 5: Comparison of programs for differential gene expression identification. Based on (Rapaport et al., 2013; Seyednasrollah et al., 2013; Schurch et al., 2015).

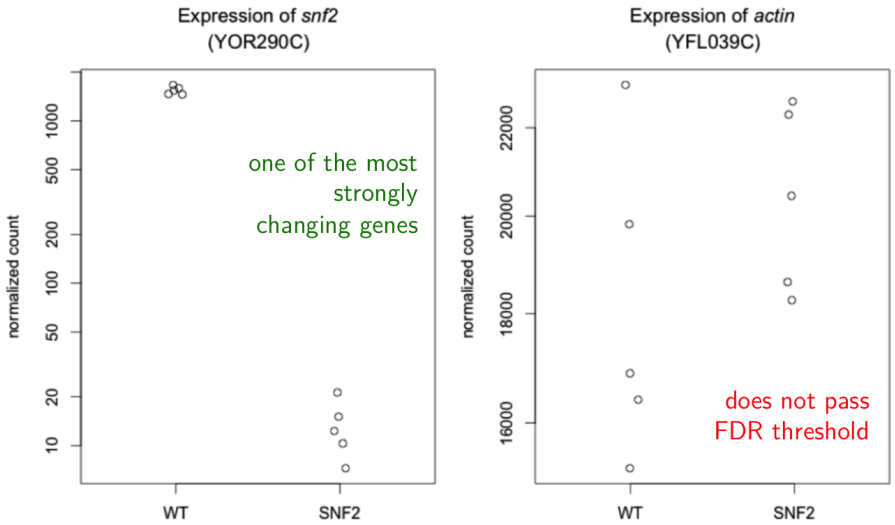
Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

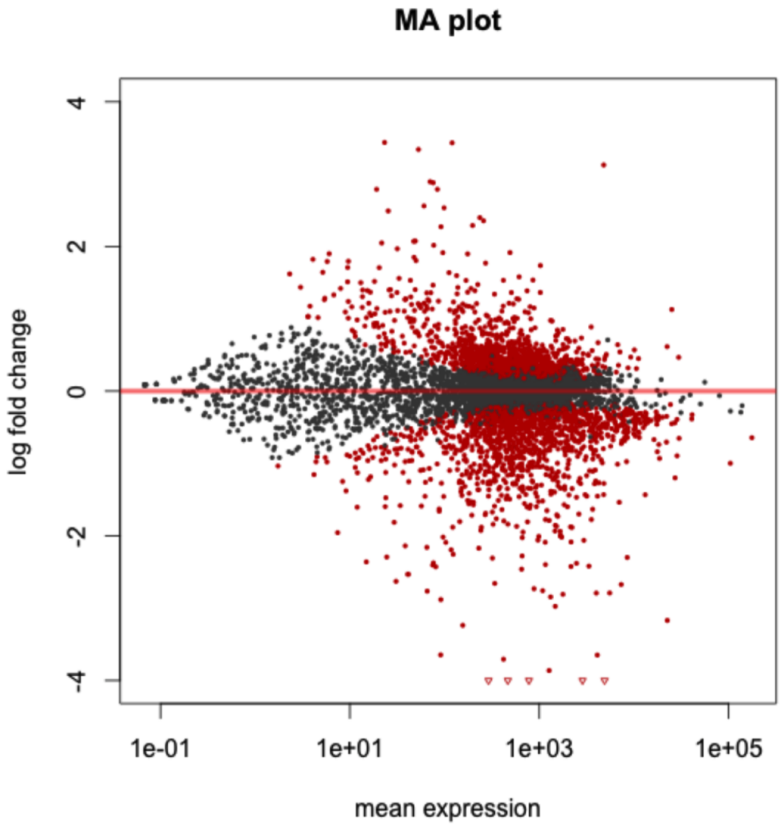
When in doubt, compare the results of limma, edgeR, and DESeq2 to get a feeling for how robust your favorite DE genes are.

Downstream analyses

Understanding the RESULTS of the DGE analysis

- Investigate the `results()` output:
  - How many DE genes? (FDR/q-value!)
  - How strongly do the DE genes change?
  - Directions of change?
  - Are your favorite genes among the DE genes?





Understanding the FUNCTIONS of your DE genes

There are myriad tools for this – many are web-based, many are R packages, many will address very specific questions. Typical points of interest are:

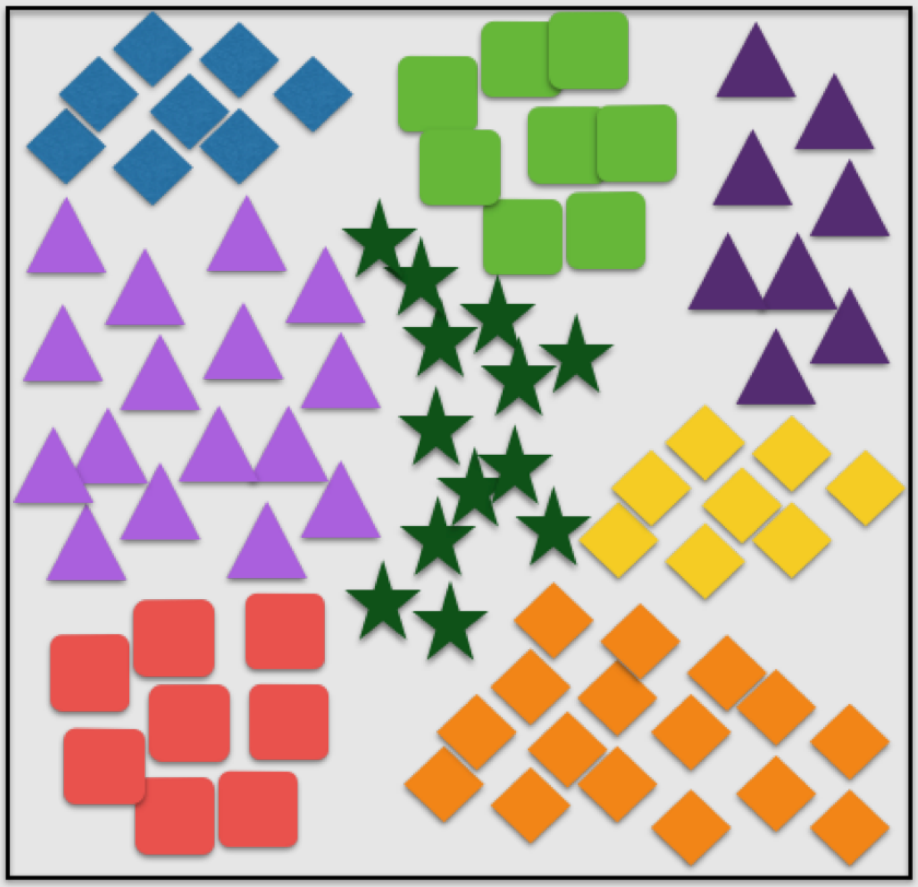
- enriched gene ontology (GO) terms
  - ontology = standardized vocabulary
  - 3 classes of gene ontologies are maintained:
    - biological processes (BP), cell components (CC), and molecular functions (MF)
- enriched pathways
  - gene sets: e.g. from MSigDB [Liberzon et al., 2015]
  - physical interaction networks: e.g. from STRING [Szklarczyk et al., 2017]
  - metabolic (and other) pathways: e.g. from KEGG [Kanehisa et al., 2017]
- upstream regulators

None (!) of these methods should lead you to make definitive claims about the role of certain pathways for your phenotype. These are **hypothesis-generating** tools!

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)

All known genes in a species
(categorized into groups)



DEGs

HBC Training

Cate-gory	Back-ground	DE list	Over-repre-sented?
A	35/6600	25/500	likely
B	56/6600	2/500	unlikely
C	10/6600	9/500	likely

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)

- "2x2 table method"
- assessing overlap of DE genes with genes of a given pathway
- statistical test: e.g. hypergeometric test
- limitations:
  - direction of change is ignored
  - magnitude of change is ignored
  - interprets genes as well as pathways as independent entities

See Khatri et al. [2012] for details!

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)

Table S1. ORA pathway analysis tools.

Khatri et al. (2012). doi: 0.1371/journal.pcbi.1002375

Name	Scope of Analysis	P-value	Correction for Multiple Hypotheses	Availability
Onto-Express	GO	Hypergeometric, binomial, chi-square	FDR, Bonferroni, Sidak, Holm	Web
GenMAPP/ MAPPFinder	GO, KEGG, MAPP	Percentage/z-score	None	Standalone
(High throughput) GoMiner	GO	Relative enrichment, Hypergeometric	None	Standalone, Web
FatiGO	GO, KEGG	Hypergeometric	None	Web
GOstat	GO	Chi-square	FDR	
GOTree Machine	GO	Hypergeometric	None	Web
FuncAssociate	GO	Hypergeometric	Bootstrap	Web
GOToolBox	GO	Hypergeometric	Bonferroni, Holm, FDR, Hommel, Hochberg	
GeneMerge	GO	Hypergeometric	Bonferroni	Web
GOEAST	GO	Hypergeometric, Chi-square	Benjamini-Yekutieli	Web
ClueGO	GO, KEGG, BioCarta, User defined	Hypergeometric	Bonferroni, Bonferroni step-down, Benjamini-Hochberg	Standalone

Two typical approaches of enrichment analyses

2. Functional Class Scoring ("Gene set enrichment")

- gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic
- score will depend on size of the pathway, and the amount of correlation between genes in the pathway
- all genes are used
- direction and magnitude of change matter
- coordinated changes of genes within the same pathway matter, too

Two typical approaches of enrichment analyses

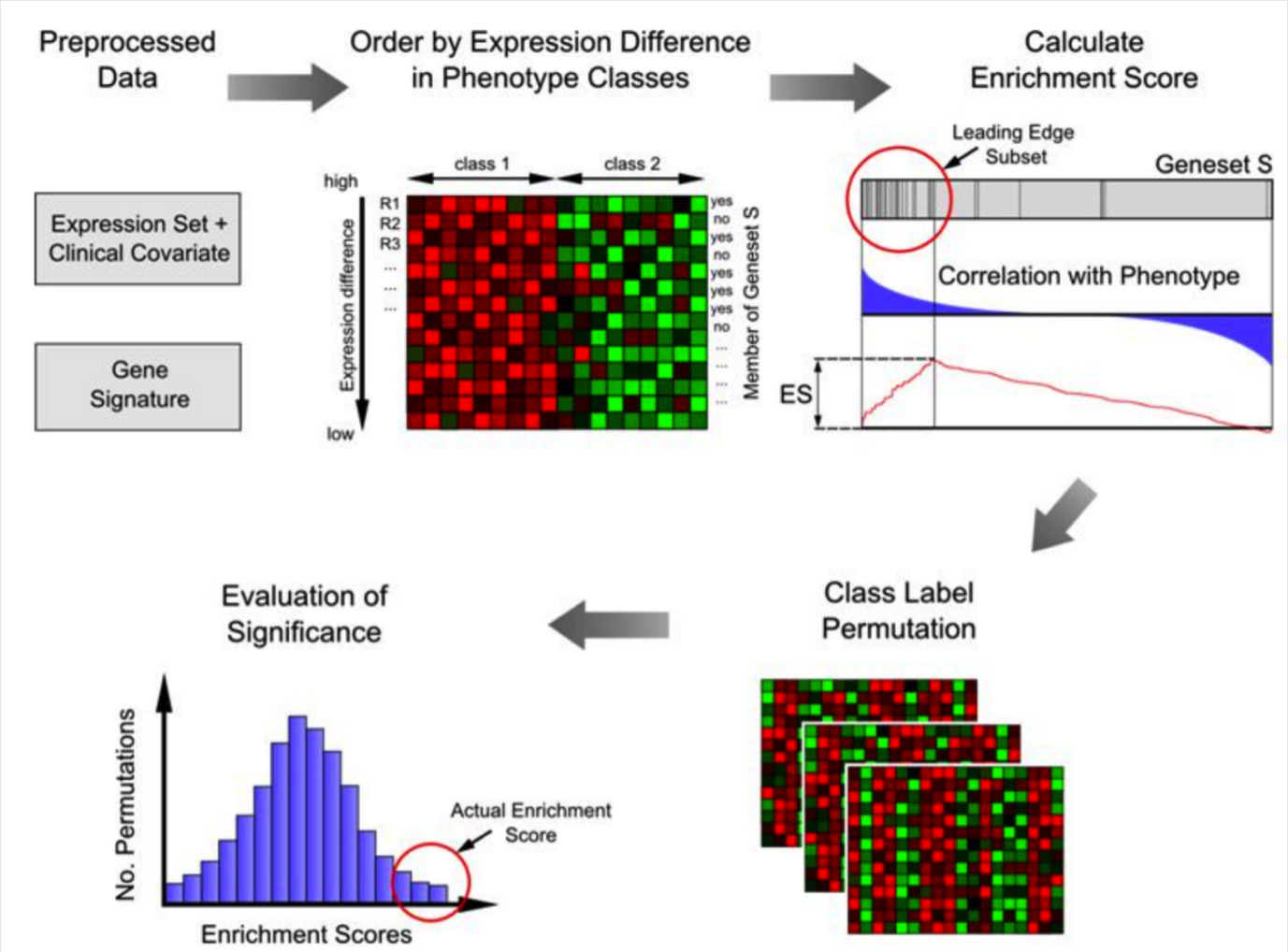
2. Functional Class Scoring ("Gene set enrichment")

Table S2. FCS pathway analysis tools.
Khatri et al. (2012). doi: 0.1371/journal.pcbi.1002375

Name	Scope of Analysis	Gene-level Statistic	Gene Set Statistic	P-value	Correction for Multiple Hypotheses	Availability
GSEA	GO, KEGG, BioCarta, MAPP, transcription factors, microRNA, cancer molecules	Signal-to-noise ratio, t-test, cosine, euclidian and manhattan distance, Pearson correlation, (log2) fold-change, log difference	Kolmogorov-Smirnov	Phenotype permutation, Gene set permutation	FDR	Standalone, R package
sigPathway	GO, KEGG, BioCarta, humanpaths	t-statistic	Wilcoxon rank sum	Phenotype permutation, Gene set permutation	FDR (NPMLE)	R package
Category	GO, KEGG	t-statistic		Phenotype permutation	NA	R package
SAFE	GO, KEGG, PFAM	Student's t-test, Welch's t-test, SAM t-test, f-statistic, Cox proportional hazards model, linear regression	Wilcoxon rank sum, Fisher's exact test statistic, Pearson's test, t-test of average difference	Phenotype permutation	FWER (Bonferroni, Holm's step-up), FDR (Benjamini-Hochberg, Yekutieli-Benjamini)	R package
GlobalTest	GO, KEGG	NA	simple and multinomial logistic regression, Q-statistics mean	Phenotype permutation, asymptotic distribution, Gamma distribution	NA	R package
PCOT2	User specified	Hotelling's T^2		Phenotype permutation, gene set permutation	FDR (Benjamini-Hochberg, Yekutieli-Benjamini), FWER (Bonferroni, Holm, Hochberg, Hommel)	R package
SAM-GS	User specified	d-statistic	sum of squared d-statistic	Phenotype permutation	FDR	Excel plug-in

Two typical approaches of enrichment analyses

2. Functional Class Scoring: Example GSEA

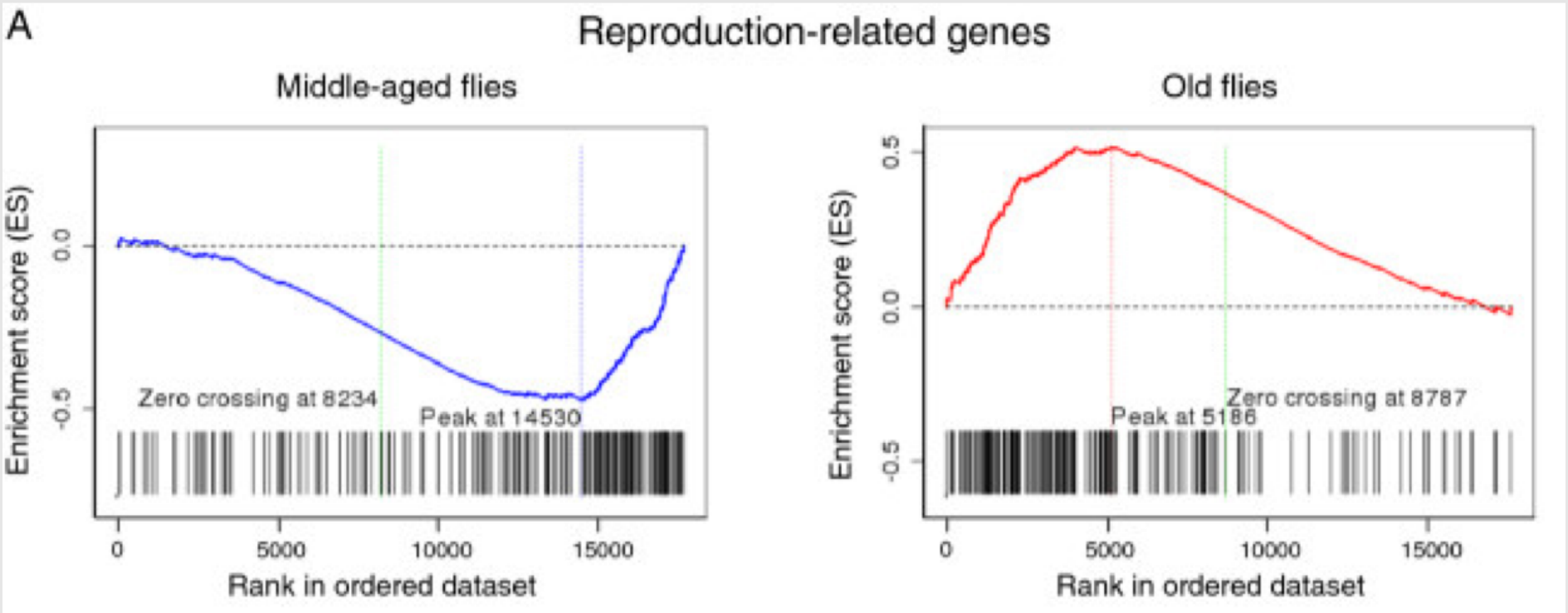


http://slideplayer.biz.tr/slide/2738467/10/images/20/Gene+Set+Enrichment+Analysis+(GSEA).jpg

Two typical approaches of enrichment analyses

2. Functional Class Scoring ("Gene set enrichment")

Example GSEA results for positive and negative correlation



Doroszuk et al. (2012) doi: 10.1186/1471-2164-13-167

Summary – downstream analyses

Know your biological question(s) of interest!

- all enrichment methods potentially suffer from gene length bias
  - long genes will get more reads [Young et al., 2010]
- for GO terms:
  - use `goseq` to identify enriched GO terms
  - use additional tools, such as GOrilla, REVIGO [Eden et al., 2009, Supek et al., 2011] to summarize the often redundant GO term lists
- for KEGG pathways:
  - e.g. GAGE and PATHVIEW [Luo and Brouwer, 2013, Luo et al., 2017] [4]
- miscellaneous including attempts to predict upstream regulators
  - Enrichr [Chen et al., 2013]
  - RegulatorTrail [Kehl et al., 2017]
  - Ingenuity Pathway Analysis Studio (proprietory software!)

See the additional links and material on our course website!

[4] https://www.r-bloggers.com/tutorial-rna-seq-differential-expression-pathway-analysis-with-sailfish-deseq2-gage-and-pathview/

References

[Anders and Huber, 2010, D'haeseleer, 2005, Dillies et al., 2013, Doroszuk et al., 2012, Subramanian et al., 2005, Dündar et al., 2018]

References

Simon Anders and Wolfgang Huber. DESeq: Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. doi: 10.1186/gb-2010-11-10-r106.

Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V. Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 2013. doi: 10.1186/1471-2105-14-128. URL http://amp.pharm.mssm.edu/Enrichr.

Patrik D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501, 2005. doi: 10.1038/nbt1205-1499.

Marie Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Nicolas Servant Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013. doi: 10.1093/bib/bbs046.

Agnieszka Doroszuk, Martijs J. Jonker, Nicolien Pul, Timo M. Breit, and Bas J. Zwaan. Transcriptome analysis of a long-lived natural Drosophila variant: a prominent role of stress- and reproduction-genes in lifespan extension. *BMC Genomics*, 2012. doi: 10.1186/1471-2164-13-167.

Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq, 2018. URL http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf.

Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, jan 2009. doi: 10.1186/1471-2105-10-48. URL http://cbl-gorilla.cs.technion.ac.il.

R. Irizarry and M. Love. Leanpub, 2015. URL https://leanpub.com/dataanalysisforthelifesciences.

R. Irizarry and M. Love. Biomedical Data Science, 2016.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkw1092.

Tim Kehl, Lara Schneider, Florian Schmidt, Daniel Stöckel, Nico Gerstner, Christina Backes, Eckart Meese, Andreas Keller, Marcel H. Schulz, and Hans Peter Lenhof. RegulatorTrail: A web service for the identification of key transcriptional regulators. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkx350. URL https://regulatortrail.bioinf.uni-sb.de/.

Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 2012. doi: 10.1371/journal.pcbi.1002375.

William B. King. Model Formulae Tutorial, 2016. URL http://ww2.coastal.edu/kingw/statistics/R-tutorials/formulae.html.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 2015. doi: 10.1016/j.cels.2015.12.004.

Michael I Love, Charlotte Soneson, and Rob Patro. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research*, 7(952), 2018. doi: 10.12688/f1000research.15398.1.

Weijun Luo and Cory Brouwer. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 2013. doi: 10.1093/bioinformatics/btt285.

Weijun Luo, Gaurav Pant, Yeshvant K. Bhavnasi, Steven G. Blanchard, and Cory Brouwer. Pathview Web: User friendly pathway visualization and data integration. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkx372. URL https://pathview.uncc.edu/.

Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4(0):1521, 2015. doi: 10.12688/f1000research.7563.2.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, Oct 2005. doi: 10.1073/pnas.0506580102. URL http://software.broadinstitute.org/gsea/index.jsp.

Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6 (7):e21800, jan 2011. doi: 10.1371/journal.pone.0021800. URL http://revigo.irb.hr/.

Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian Von Mering. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkw937.

Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 2010. doi: 10.1186/gb-2010-11-2-r14.